Modeling Reference Interviews as a Basis for Improving Automatic QA Systems

Nancy J. McCracken, Anne R. Diekema, Grant Ingersoll, Sarah C. Harwell, Eileen E. Allen, Ozgur Yilmazel, Elizabeth D. Liddy
Center for Natural Language Processing
Syracuse University
Syracuse, NY 13244
{ njmccrac, diekemar, gsingers, scharwel, eeallen, oyilmaz, liddy}@syr.edu

Abstract

The automatic QA system described in this paper uses a reference interview model to allow the user to guide and contribute to the QA process. A set of system capabilities was designed and implemented that defines how the user's contributions can help improve the system. These include tools, called the Query Template Builder and the Knowledge Base Builder, that tailor the document processing and QA system to a particular domain by allowing a Subject Matter Expert to contribute to the query representation and to the domain knowledge. During the QA process, the system can interact with the user to improve query terminology by using Spell Checking, Answer Type verification, Expansions and Acronym Clarifications. The system also has capabilities that depend upon, and expand the user's history of interaction with the system, including a User Profile, Reference Resolution, and Question Similarity modules

1 Introduction

Reference librarians have successfully fielded questions of all types for years using the Reference Interview to clarify an unfocused question, narrow a broad question, and suggest further information that the user might not have thought to ask for. The reference interview tries to elicit sufficient information about the user's real need to enable a librarian to understand the question enough to begin searching. The question is clarified, made more specific, and contextualized with relevant detail. Real questions from real users are often "ill-formed' with respect to the information system; that is, they do not match the structure of 'expectations' of the system (Ross et al., 2002). A reference interview translates the user's question into a representation that the librarian and the library systems can interpret correctly. The human reference interview process provides an ideal, well-tested model of how questioner and answerer work together co-operatively and, we believe, can be successfully applied to the digital environment. The findings of researchers applying this model in online situations (Bates, 1989, Straw, 2004) have enabled us to understand how a system might work with the user to provide accurate and relevant answers to complex questions.

Our long term goal in developing Question-Answering (QA) systems for various user groups is to permit, and encourage users to positively contribute to the QA process, to more nearly mirror what occurs in the reference interview, and to develop an automatic QA system that provides fuller, more appropriate, individually tailored responses than has been available to date.

Building on our Natural Language Processing (NLP) experience in a range of information access applications, we have focused our QA work in two areas: 1) modeling the subject domain of the collections of interest to a set of

Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006, pages 17–24,
New York City, NY, USA. June 2006.

users for whom we are developing the QA system, and; 2) modeling the query clarification and negotiation interaction between the information seeker and the information provider. Examples of these implementation environments are:

1. Undergraduate aerospace engineering students working in collaborative teams on course projects designing reusable launch vehicles, who use a QA system in their course-related research.
2. Customers of online business sites who use a QA system to learn more about the products or services provided by the company, or who wish to resolve issues concerning products or service delivery.

In this paper, we describe the capabilities we have developed for these specific projects in order to explicate a more general picture of how we model and utilize both the domains of inquiry and typical interaction processes observed in these diverse user groups.

2 Background and related research

Our work in this paper is based on two premises: 1) user questions and responsive answers need to be understood within a larger model of the user's information needs and requirements, and, 2) a good interactive QA system facilitates a dialogue with its users to ensure it understands and satisfies these information needs. The first premise is based on the long-tested and successful model of the reference interview (Bates, 1997, Straw, 2004), which was again validated by the findings of an ARDA-sponsored workshop to increase the research community's understanding of the information seeking needs and cognitive processes of intelligence analysts (Liddy, 2003). The second premise instantiates this model within the digital and distributed information environment.

Interactive QA assumes an interaction between the human and the computer, typically through a combination of a clarification dialogue and user modeling to capture previous interactions of users with the system. De Boni et al. (2005) view the clarification dialogue mainly as the presence or absence of a relationship between the question from the user and the answer provided by the system. For example, a user may ask a question, receive an answer and ask another question in order to clarify the meaning, or, the user may ask an additional question which expands on the previous answer. In their research De Boni et al. (2005) try to determine automatically whether or not there exists a relationship between a current question and preceding questions, and if there is a relationship, they use this additional information in order to determine the correct answer.

We prefer to view the clarification dialogue as more two-sided, where the system and the user actually enter a dialogue, similar to the reference interview as carried out by reference librarians (Diekema et al., 2004). The traditional reference interview is a cyclical process in which the questioner poses their question, the librarian (or the system) questions the questioner, then locates the answer based on information provided by the questioner, and returns an answer to the user who then determines whether this has satisfied their information need or whether further clarification or further questions are needed. The HITIQA system's (Small et al., 2004) view of a clarification system is closely related to ours—their dialogue aligns the understanding of the question between system and user. Their research describes three types of dialogue strategies: 1) narrowing the dialogue, 2) broadening the dialogue, and 3) a fact seeking dialogue.

Similar research was carried out by Hori et al. (2003), although their *system* automatically determines whether there is a need for a dialogue, not the *user*. The system identifies ambiguous questions (i.e. questions to which the system could not find an answer). By gathering additional information, the researchers believe that the system can find answers to these questions. Clarifying questions are automatically generated based on the ambiguous question to solicit additional information from the user. This process is completely automated and based on templates that generate the questions. Still, removing the cognitive burden from the user through automation is not easy to implement and can be the cause of error or misunderstanding. Increasing user involvement may help to reduce this error.

As described above, it can be seen that interactive QA systems have various levels of dialogue automation ranging from fully automatic (De Boni et al., 2004, Hori et al., 2004) to a strong

user involvement (Small et al., 2004, Diekema et al., 2004). Some research suggests that clarification dialogues in open-domain systems are more unpredictable than those in restricted domain systems, the latter lending itself better to automation (Hori et al., 2003, Jönsson et al., 2004). Incorporating the user's inherent knowledge of the intention of their query is quite feasible in restricted domain systems and should improve the quality of answers returned, and make the experience of the user a less frustrating one. While many of the systems described above are promising in terms of IQA, we believe that incorporating knowledge of the user in the question negotiation dialogue is key to developing a more accurate and satisfying QA system.

3 System Capabilities

In order to increase the contribution of users to our question answering system, we expanded our traditional domain independent QA system by adding new capabilities that support system-user interaction.

3.1 Domain Independent QA

Our traditional domain-independent QA capability functions in two stages, the first information retrieval stage selecting a set of candidate documents, the second stage doing the answer finding within the filtered set. The answer finding process draws on models of question types and document-based knowledge to seek answers without additional feedback from the user. Again, drawing on the modeling of questions as they interact with the domain representation, the system returns answers of variable lengths on the fly in response to the nature of the question since factoid questions may be answered with a short answer, but complex questions often require longer answers. In addition, since our QA projects were based on closed collections, and since closed collections may not provide enough redundancy to allow for short answers to be returned, the variable length answer capability assists in finding answers to factoid questions. The QA system provides answers in the form of short answers, sentences, and answer-providing passages, as well as links to the full answer-providing documents. The user can provide relevance feedback by selecting the full documents that offer the best information. Using this feedback, the system can reformulate the question and look for a better set of documents from which to find an answer to the question. Multiple answers can be returned, giving the user a more complete picture of the information held within the collection.

One of our first tactics to assist in both question and domain modeling for specific user needs was to develop tools for Subject Matter Experts (SMEs) to tailor our QA systems to a particular domain. Of particular interest to the interactive QA community is the Query Template Builder (QTB) and the Knowledge Base Builder (KBB).

Both tools allow a priori alterations to question and domain modeling for a community, but are not sensitive to *particular* users. Then the interactive QA system permits question- and user-specific tailoring of system behavior simply because it allows subject matter experts to change the way the system understands their need at the time of the search.

Question Template Builder (QTB) allows a subject matter expert to fine tune a question representation by adding or removing stopwords on a question-by-question basis, adding or masking expansions, or changing the answer focus. The QTB displays a list of Question-Answer types, allows the addition of new Answer Types, and allows users to select the expected answer type for specific questions. For example, the subject matter expert may want to adjust particular "who" questions as to whether the expected answer type is "person" or "organization". The QTB enables organizations to identify questions for which they want human intervention and to build specialized term expansion sets for terms in the collection. They can also adjust the stop word list, and refine and build the Frequently or Previously Asked Question (FAQ/PAQ) collection.

Knowledge Base Builder (KBB) is a suite of tools developed for both commercial and government customers. It allows the users to view and extract terminology that resides in their document collections. It provides useful statistics about the corpus that may indicate portions that require attention in customization. It collects frequent / important terms with categorizations to enable ontology building (semi-automatic, permitting human review), term collocation for use

in identifying which sense of a word is used in the collection for use in term expansion and categorization review. KBB allows companies to tailor the QA system to the domain vocabulary and important concept types for their market. Users are able to customize their QA applications through human-assisted automatic procedures. The Knowledge Bases built with the tools are

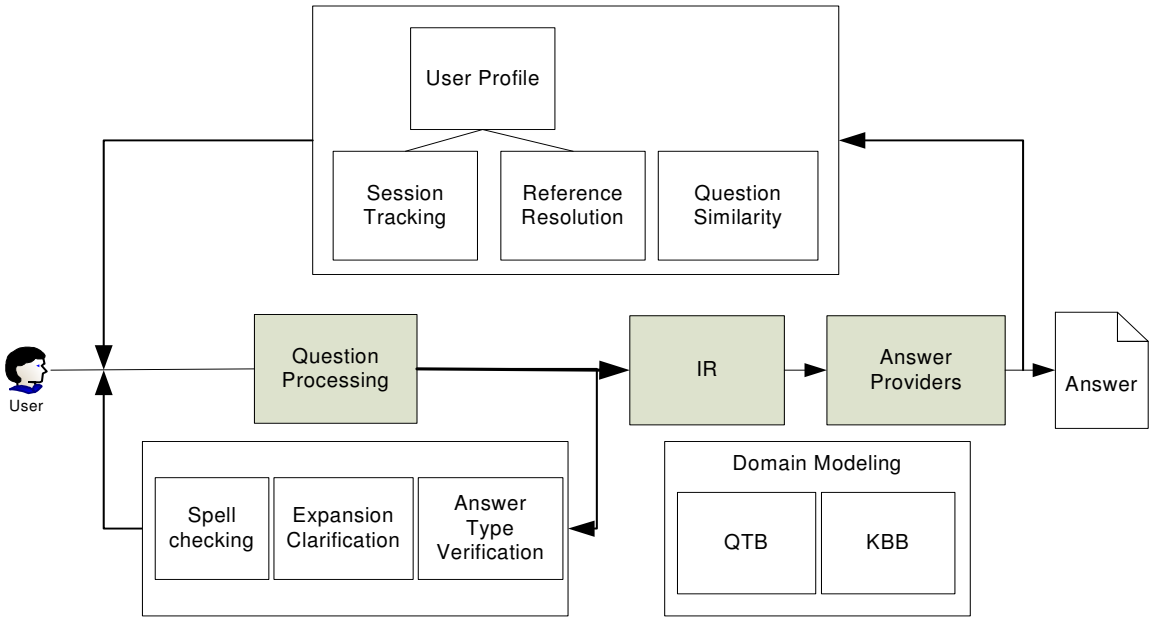

Figure 1. System overview

primarily lexical semantic taxonomic resources. These are used by the system in creating frame representations of the text. Using automatically harvested data, customers can review and alter categorization of names and entities and expand the underlying category taxonomy to the domain of interest. For example, in the NASA QA system, experts added categories like "material", "fuel", "spacecraft" and "RLV", (Reusable Launch Vehicles). They also could specify that "RLV" is a subcategory of "spacecraft" and that space shuttles like "Atlantis" have category "RLV". The KBB works in tandem with the QTB, where the user can find terms in either documents or example queries

3.2 Interactive QA Development

In our current NASA phase, developed for undergraduate aerospace engineering students to quickly find information in the course of their studies on reusable launch vehicles, the user can view immediate results, thus bypassing the Reference Interviewer, or they may take the opportunity to utilize its increased functionality and interact with the QA system. The capabilities we have developed, represented by modules added to the system, fall into two groups. Group One includes capabilities that draw on direct interaction with the user to clarify what is being asked and that address terminological issues. It includes Spell Checking, Expansion Clarification, and Answer Type Verification. Answers change dynamically as the user provides more input about what was meant. Group Two capabilities are dependent upon, and expand upon the user's history of interaction with the system and include User Profile, Session Tracking, Reference Resolution, Question Similarity and User Frustration Recognition modules. These gather knowledge about the user, help provide co-reference resolution within an extended dialogue, and monitor the level of frustration a user is experiencing.

The capabilities are explained in greater detail below. Figure 1 captures the NASA system process and flow.

Group One:

In this group of interactive capabilities, after the user asks a query, answers are returned as in a typical system. If the answers presented aren't satisfactory, the system will embark on a series of interactive steps (described below) in which alternative spelling, answertypes, clarifications and expansions will be suggested. The user can choose from the system's suggestions or type in their own. The system will then revise the query and return a new set of answers. If those answers aren't satisfactory, the user can continue interacting with the system until appropriate answers are found.

Spell checking: Terms not found in the index of the document collection are displayed as potentially misspelled words. In this preliminary phase, spelling is checked and users have the opportunity to select correct and/or alternative spellings.

AnswerType verification: The interactive QA system displays the type of answer that the system is looking for in order to answer the question. For example for the question, *Who piloted the first space shuttle?,* the answer type is 'person', and the system will limit the search for candidate short answers in the collection to those that are a person's name. The user can either accept the system's understanding of the question or reject the type it suggests. This is particularly useful in semantically ambiguous questions such as "Who makes Mountain Dew?" where the system might interpret the question as needing a person, but the questioner actually wants the name of a company.

Expansion: This capability allows users to review the possible relevant terms (synonyms and group members) that could enhance the question-answering process. The user can either select or deselect terms of interest which do or do not express the intent of the question. For example, if the user asks: *How will aerobraking change the orbit size?* then the system can bring back the following expansions for "aerobraking": *By aerobraking do you mean the following: 1) aeroassist, 2) aerocapture, 3) aeromaneuvering, 4) interplanetary transfer orbits, or 5) transfer orbits.*

Acronym Clarification: For abbreviations or acronyms within a query, the full explications known by the system for the term can be displayed back to the user. The clarifications implemented are a priori limited to those that are relevant to the domain. In the aerospace domain for example, if the question was *What is used for the TPS of the RLV?,* the clarifications of TPS would be *thermal protection system*, *thermal protection subsystem*, *test preparation sheet*, or *twisted pair shielded*, and the clarification of RLV would be *reusable launch vehicle*. The appropriate clarifications can be selected to assist in improving the search. For a more generic domain, the system would offer broader choices. For example, if the user types in the question: *What educational programs does the AIAA offer?,* then the system might return: *By AIAA, do you mean (a) American Institute of Aeronautics and Astronautics (b) Australia Indonesia Arts Alliance or (c) Americans for International Aid & Adoption?*

Group Two:

User Profile: The User Profile keeps track of more permanent information about the user. The profile includes a small standard set of user attributes, such as the user's name and / or research interests. In our commercially funded work, selected information gleaned from the question about the user was also captured in the profile. For example, if a user asks "How much protein should my husband be getting every day?", the fact that the user is married can be added to their profile for future marketing, or for a new line of dialogue to ask his name or age. This information is then made available as context information for the QA system to resolve references that the user makes to themselves and their own attributes.

For the NASA question-answering capability, to assist students in organizing their questions and results, there is an area for users to save their searches as standing queries, along with the results of searching (Davidson, 2006). This information, representing topics and areas of interest, can help to focus answer finding for new questions the user asks.

Not yet implemented, but of interest, is the ability to save information such as a user's

preferences (format, reliability, sources), that could be used as filters in the answer finding process.

Reference Resolution: A basic feature of an interactive QA system is the requirement to understand the user's questions and responsive answers as one session. The sequence of questions and answers forms a natural language dialogue between the user and the system. This necessitates NLP processing at the discourse level, a primary task of which is to resolve references across the session. Building on previous work in this area done for the Context Track of TREC 2001 (Harabagiu et al, 2001) and additional work (Chai and Jin, 2004) suggesting discourse structures are needed to understand the question/answer sequence, we have developed session-based reference resolution capability. In a dialogue, the user naturally includes referring phrases that require several types of resolution.

The simplest case is that of referring pronouns, where the user is asking a follow-up question, for example:

Q1: When did Madonna enter the music business?
A1: Madonna's first album, Madonna, came out in 1983 and since then she's had a string of hits, been a major influence in the music industry and become an international icon.
Q2: When did she first move to NYC?

In this question sequence, the second question contains a pronoun, "she", that refers to the person "Madonna" mentioned both in the previous question and its answer. Reference resolution would transform the question into "When did Madonna first move to NYC?"

Another type of referring phrase is the definite common noun phrase, as seen in the next example:

Q1: If my doctor wants me to take Acyclovir, is it expensive?
A1: Glaxo-Wellcome, Inc., the company that makes Acyclovir, has a program to assist individuals that have HIV and Herpes.
Q2: Does this company have other assistance programs?

The second question has a definite noun phrase "this company" that refers to "Glaxo-Wellcome, Inc." in the previous answer, thus transforming the question to "Does Glaxo-Wellcome, Inc. have other assistance programs?"

Currently, we capture a log of the question/answer interaction, and the reference resolution capability will resolve any references in the current question that it can by using linguistic techniques on the discourse of the current session. This is almost the same as the narrative coreference resolution used in documents, with the addition of the need to understand first and second person pronouns from the dialogue context. The coreference resolution algorithm is based on standard linguistic discourse processing techniques where referring phrases and candidate resolvents are analyzed along a set of features that typically includes gender, animacy, number, person and the distance between the referring phrase and the candidate resolvent.

Question Similarity: Question Similarity is the task of identifying when two or more questions are related. Previous studies (Boydell et al., 2005, Balfe and Smyth, 2005) on information retrieval have shown that using previously asked questions to enhance the current question is often useful for improving results among like-minded users. Identifying related questions is useful for finding matches to Frequently Asked Questions (FAQs) and Previously Asked Questions (PAQs) as well as detecting when a user is failing to find adequate answers and may be getting frustrated. Furthermore, similar questions can be used during the reference interview process to present questions that other users with similar information needs have used and any answers that they considered useful.

CNLP's question similarity capability comprises a suite of algorithms designed to identify when two or more questions are related. The system works by analyzing each query using our Language-to-Logic (L2L) module to identify and weight keywords in the query, provide expansions and clarifications, as well as determine the focus of the question and the type of answer the user is expecting (Liddy et al., 2003). We then compute a series of similarity measures on two or more L2L queries. Our measures adopt a variety of approaches, including those that are based on keywords in the query: cosine similarity, keyword string matching, expansion analysis, and spelling variations. In addition, two measures are based on the representation of the whole query:answer type

and answer frame analysis. An answer frame is our representation of the meaningful extractions contained in the query, along with metadata about where they occur and any other extractions that relate to in the query.

Our system will then combine the weighted scores of two or more of these measures to determine a composite score for the two queries, giving more weight to a measure that testing has determined to be more useful for a particular task.

We have utilized our question similarity module for two main tasks. For FAQ/PAQ (call it XAQ) matching, we use question similarity to compare the incoming question with our database of XAQs. Through empirical testing, we determined a threshold above which we consider two questions to be similar.

Our other use of question similarity is in the area of frustration detection. The goal of frustration detection is to identify the signs a user may be giving that they are not finding relevant answers so that the system can intervene and offer alternatives before the user leaves the system, such as similar questions from other users that have been successful.

4 Implementations:

The refinements to our Question Answering system and the addition of interactive elements have been implemented in three different, but related working systems, one of which is strictly an enhanced IR system. None of the three incorporates all of these capabilities. In our work for MySentient, Ltd, we developed the session-based reference resolution capability, implemented the variable length and multiple answer capability, modified our processing to facilitate the building of a user profile, added FAQ/PAQ capability, and our Question Similarity capability for both FAQ/PAQ matching and frustration detection. A related project, funded by Syracuse Research Corporation, extended the user tools capability to include a User Interface for the KBB and basic processing technology. Our NASA project has seen several phases. As the project progressed, we added the relevant developed capabilities for improved performance. In the current phase, we are implementing the capabilities which draw on user choice.

5 Conclusions and Future Work

The reference interview has been implemented as an interactive dialogue between the system and the user, and the full system is near completion. We are currently working on two types of evaluation of our interactive QA capabilities. One is a system-based evaluation in the form of unit tests, the other is a user-based evaluation. The unit tests are designed to verify whether each module is working correctly and whether any changes to the system adversely affect results or performance. Crafting unit tests for complex questions has proved challenging, as no gold standard for this type of question has yet been created. As the data becomes available, this type of evaluation will be ongoing and part of regular system development.

As appropriate for this evolutionary work within specific domains for which there are not gold standard test sets, our evaluation of the QA systems has focused on qualitative assessments. What has been a particularly interesting outcome is what we have learned in elicitation from graduate students using the NASA QA system, namely that they have multiple dimensions on which they evaluate a QA system, not just traditional recall and precision (Liddy et al, 2004). The high level dimensions identified include system performance, answers, database content, display, and expectations. Therefore the evaluation criteria we believe appropriate for IQA systems are centered around the display (UI) category as described in Liddy et al, (2004). We will evaluate aspects of the UI input subcategory, including question understanding, information need understanding, querying style, and question formulation assistance. Based on this user evaluation the system will be improved and retested.

References

Evelyn Balfe and Barry Smyth. 2005. An Analysis of Query Similarity in Collaborative Web Search. In *Proceedings of the 27th European Conference on Information Retrieval*. Santiago de Compostela, Spain.

Marcia J. Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13: 407-424.

Mary Ellen Bates. 1997. The Art of the Reference Interview. *Online World*. September 15.

Oisín Boydell, Barry Smyth, Cathal Gurrin, and Alan F. Smeaton. 2005. A Study of Selection Noise in Collaborative Web Search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland. http://www.ijcai.org/papers/post-0214.pdf

Joyce Y. Chai, and Rong Jin. 2004. Discourse Structure for Context Question Answering. In *Proceedings of the Workshp on the Pragmatics of Quesiton Answering,* HST-NAACL, Boston. http://www.cse.msu.edu/~rongjin/publications/HLTQAWorkshop04.pdf

Barry D. Davidson. 2006. *An Advanced Interactive Discovery Learning Environment for Engineering Education: Final Report.* Submitted to R. E. Gillian, National Aeronautics and Space Administration.

Marco De Boni and Suresh Manandhar. 2005. Implementing Clarification Dialogues in Open Domain Question Answering. *Natural Language Engineering* 11(4): 343-361.

Anne R. Diekema, Ozgur Yilmazel, Jiangping Chen, Sarah Harwell, Lan He, and Elizabeth D. Liddy. 2004. Finding Answers to Complex Questions. In *New Directions in Question Answering*. (Ed.) Mark T. Maybury. The MIT Press, 141-152.

Sanda Harabagiu, Dan Moldovan, Marius Paşca, Mihai Surdeanu, Rada Mihalcea, Roxana Gîrju, Vasile Rus, Finley Lăcătuşu, Paul Morărescu, Răzvan Bunescu. 2001. Answering Complex, List and Context Questions with LCC's Question-Answering Server, TREC 2001.

Chiori Hori, Takaaki Hori., Hideki Isozaki, Eisaku Maeda, Shigeru Katagiri, and Sadaoki Furui. 2003. Deriving Disambiguous Queries in a Spoken Interactive ODQA System. In *ICASSP*. Hongkong, I: 624-627.

Arne Jönsson, Frida Andén, Lars Degerstedt, Annika Flycht-Eriksson, Magnus Merkel, and Sara Norberg. 2004. Experiences from Combining Dialogue System Development With Information Extraction Techniques. In *New Directions in Question Answering*. (Ed.) Mark T. Maybury. The MIT Press, 153-164.

Elizabeth D. Liddy. 2003. Question Answering in Contexts. Invited Keynote Speaker. ARDA AQUAINT Annual Meeting. Washington, DC. Dec 2-5, 2003.

Elizabeth D. Liddy, Anne R. Diekema, Jiangping Chen, Sarah Harwell, Ozgur Yilmazel, and Lan He. 2003. What do You Mean? Finding Answers to Complex Questions. *Proceedings of New Directions in Question Answering*. AAAI Spring Symposium, March 24-26.

Elizabeth D. Liddy, Anne R. Diekema, and Ozgur Yilmazel. 2004. Context-Based Question-Answering Evaluation. In *Proceedings of the 27th Annual ACM-SIGIR Conference*. Sheffield, England

Catherine S. Ross, Kirsti Nilsen, and Patricia Dewdney. 2002. *Conducting the Reference Interview*. Neal-Schuman, New York, NY.

Sharon Small, Tomek Strzalkowski, Ting Liu, Nobuyuki Shimizu, and Boris Yamrom. 2004. A Data Driven Approach to Interactive QA. In *New Directions in Question Answering*. (Ed.) Mark T. Maybury. The MIT Press, 129-140.

Joseph E. Straw. 2004. Expecting the Stars but Getting the Moon: Negotiating around Patron Expectations in the Digital Reference Environment. In *The Virtual Reference Experience: Integrating Theory into Practice.* Eds. R. David Lankes, Joseph Janes, Linda C. Smith, and Christina M. Finneran. Neal-Schuman, New York, NY.